

# Creating a Healthcare Knowledge Graph From Statistical Open Data

## HEALTHCARE AND LIFE SCIENCES SYMPOSIUM

Enayat Rajabi

[enayat\\_rajabi@cbu.ca](mailto:enayat_rajabi@cbu.ca)

May 2nd, 2022

# Presenter

- Assistant Professor of Data Analytics, Cape Breton University, NS, Canada
- Adjunct Professor at the Faculty of Computer Science, Dalhousie University, Canada
- Canada NSERC grant holder



# Open Statistical Data

- Publishing open statistical data is getting increased attention on the Web at various levels of governments.
- Open government data increases government transparency, and accountability, contributes to economic growth and improves administrative processes.
- A variety of open statistical data are published in various domains, including public health.
- The lack of meaning behind the statistical data makes it impossible to form a network and link this kind of data to infer, create and query knowledge.

# A Knowledge Graph for Open Statistical Data

- Interconnectivity between isolated open datasets gives a machine much information to work with, thereby strengthening its ability to deduce relations and infer meaning.
- The design of a knowledge graph allows machines to draw new inferences that enrich the information available to users.
- Using classes and relationships, designing a knowledge graph links arbitrary entities or concepts from different domains.

# Canada Open Data

- Overall, Canada has 11 provinces and territories with approximately 11,771 published datasets in different domains ranging from “Business and Economy” to “Health and Wellness”.
- Each province publishes open statistical datasets via a provincial open data portal in different formats, including CSV, JSON, and Excel.
- A few of them allow users to export data in RDF format, they do not follow the Linked Data vocabulary five-star standards, and the open data portals are usually not linked to one another.

# Canada Open Data (cont.)

- The datasets are isolated within or among the data portals, while many are conceptually linked.
- Statistical data regarding diseases in a province in different years should be manually analyzed to answer some important questions like: "Which viral diseases had the most number of cases in a province in 2017?".

# Provincial Open Data - Nova Scotia

The screenshot displays the Nova Scotia Open Data portal. At the top left is the Nova Scotia logo. A search bar is located at the top right. Below the logo is a navigation menu with links for Home, Catalogue, User's Guide, Developers, and Survey. Social media icons for Facebook, Twitter, and YouTube are also present, along with a Sign In button.

A large search bar is highlighted with an orange border. Below it, the search results are displayed. The first result is "Community Health Boards" under the "Health and Wellness" category. It is a Dataset updated on April 5, 2022, with 842 views. The description states: "The Community Health Boards of Nova Scotia and the areas that they are responsible as shown by their spatial distribution." It includes a "More" link and tags: "chb, community health board, health". There is also a link to "API Docs".

The second result is "[ARCHIVED] Health Statistics Mental Health 2001-2007" under the "Health and Wellness" category. It is a Dataset updated on January 6, 2020, with 656 views. The description states: "[ARCHIVED] Community Counts data is retained for archival purposes only, such as research, reference and record-keeping. This data has not been maintained or updated. Users looking for" followed by a "More" link. It includes tags: "2001, 2007, community counts, health statistics, mental health" and a link to "API Docs".

On the left side, there are filters for "Categories" and "View Types". The "Categories" filter includes: Business and Economy, Communities and Social Services, Government Administration, and Nature and Environment. The "View Types" filter includes: Calendars, Charts, and Data Lens pages.

# Provincial Open Data - Alberta

The screenshot displays the Alberta Government Open Data portal. At the top left is the Alberta Government logo. A search bar on the top right contains the text "Search all Resources". Below the logo is a navigation menu with "Resources" and "Interact" highlighted. A secondary menu shows "Open Data", "Publications", and "Documentation". The breadcrumb trail reads "HOME / OPEN GOVERNMENT / PUBLICATIONS /". Social media icons for Facebook, Twitter, LinkedIn, Email, and RSS are present. The main heading is "PUBLICATIONS" followed by "Confirmed reportable and notifiable diseases in Alberta". Two tabs are visible: "Summary" (selected) and "Detailed Information". The "DESCRIPTION" section contains text about the number of confirmed farmed herds or flocks affected by reportable and notifiable diseases in Alberta. The "UPDATED" section shows the date "March 2, 2022". The "TAGS" section includes "OCPV", "animal diseases", "notifiable diseases", and "reportable diseases". The "RESOURCES" section features a document icon and the year "2021". At the bottom, there are buttons for "MORE INFORMATION" and "DOWNLOAD", and a note that "DOWNLOADS: 14".

Alberta Government

Search all Resources

Resources Interact

Open Data Publications Documentation

HOME / OPEN GOVERNMENT / PUBLICATIONS /

f t in e r

**PUBLICATIONS**

## Confirmed reportable and notifiable diseases in Alberta

Summary Detailed Information

**DESCRIPTION**

Information on the number of confirmed farmed herds or flocks affected by reportable and notifiable diseases in Alberta. Reportable diseases are those that are a serious threat to public health or animal health and can cause serious economic, political and social impacts to the livestock industry and/or the province. These diseases require action to control or eradicate. Notifiable diseases are diseases of concern that do not require action, but should be monitored to establish prevalence or trends.

**UPDATED**

March 2, 2022

**TAGS**

OCPV animal diseases notifiable diseases reportable diseases

**RESOURCES**

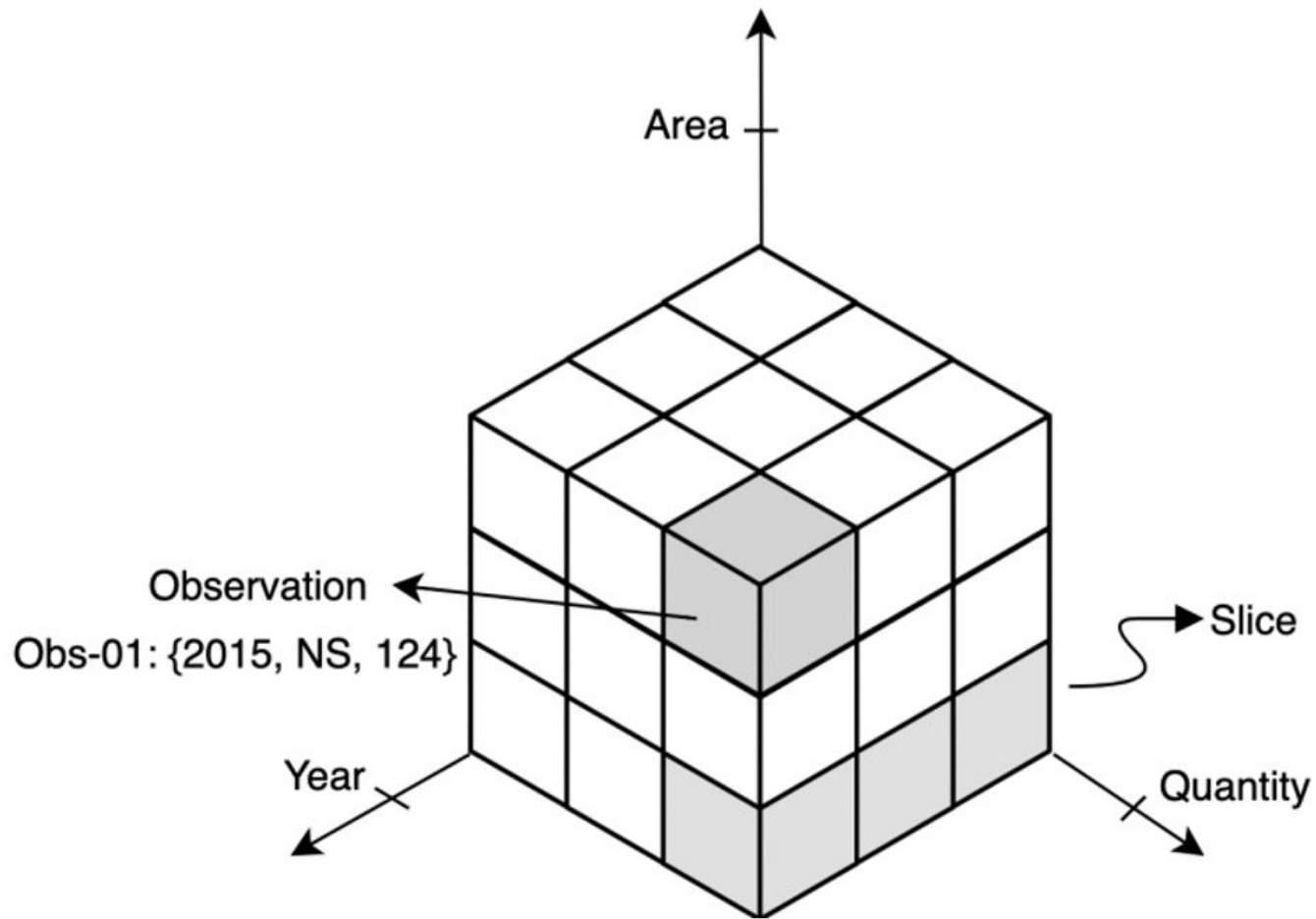
2021

MORE INFORMATION DOWNLOAD

DOWNLOADS: 14



# Data Model for Statistical Data



# Metadata in Open Statistical Data

## Notifiable Diseases Counts and Rates 2005-2017

Health And Wellness

Overall counts and rates (per 100,000 population) of notifiable diseases reported in Nova Scotia for 2005-2017.

Updated December 5, 2018

About this Dataset Mute Dataset

Updated  
**December 5, 2018**

Data Last Updated: December 5, 2018  
Metadata Last Updated: December 5, 2018

Date Created  
November 17, 2015

Views: **1,925**  
Downloads: **172**

Data Provided by: *(none)*  
Dataset Owner: Open Data Nova Scotia

### Detailed Metadata

|                        |                               |
|------------------------|-------------------------------|
| Department             | Health and Wellness           |
| Geographic Region Name | Nova Scotia                   |
| Language               | eng                           |
| Frequency              | Annually                      |
| Time Period Coverage   | 2005-01-01 through 2017-12-31 |

Usage Considerations

Current as of November 19, 2018. 1. For 2005-2008, chronic and unspecified hepatitis B were reported together.; 2. For 2009-2016, Group A Streptococcal Disease Invasive are classified as 'Severe' or 'non-Severe'; 3. Only diseases that have reported cases in the last 10 years are included.; 4. 2014: One case of Clostridium difficile and one case of MRSA did not report age group. 2015: 2 cases of MRSA and 1 case of Lyme Disease did not report age group. 2016: 1 case of MRSA and 2 cases of chlamydia did not report age group; 5. 2014: Two cases of chlamydia did not report sex. 2015: 3 cases of chlamydia did not report sex. 2016: 3 cases of chlamydia did not report sex.; 6. 2015: 5 HIV cases did not report zone. 2016: 2 HIV cases did not report zone.; 7. 2017: One case of MRSA did not report age: Two cases of Chlamydia did not report gender.

Related Documents

<http://novascotia.ca/dhw/populationhealth/>; <http://novascotia.ca/dhw/populationhealth/diseases-and-conditions-A-Z.asp>; <http://novascotia.ca/dhw/cdpc/cdc/>

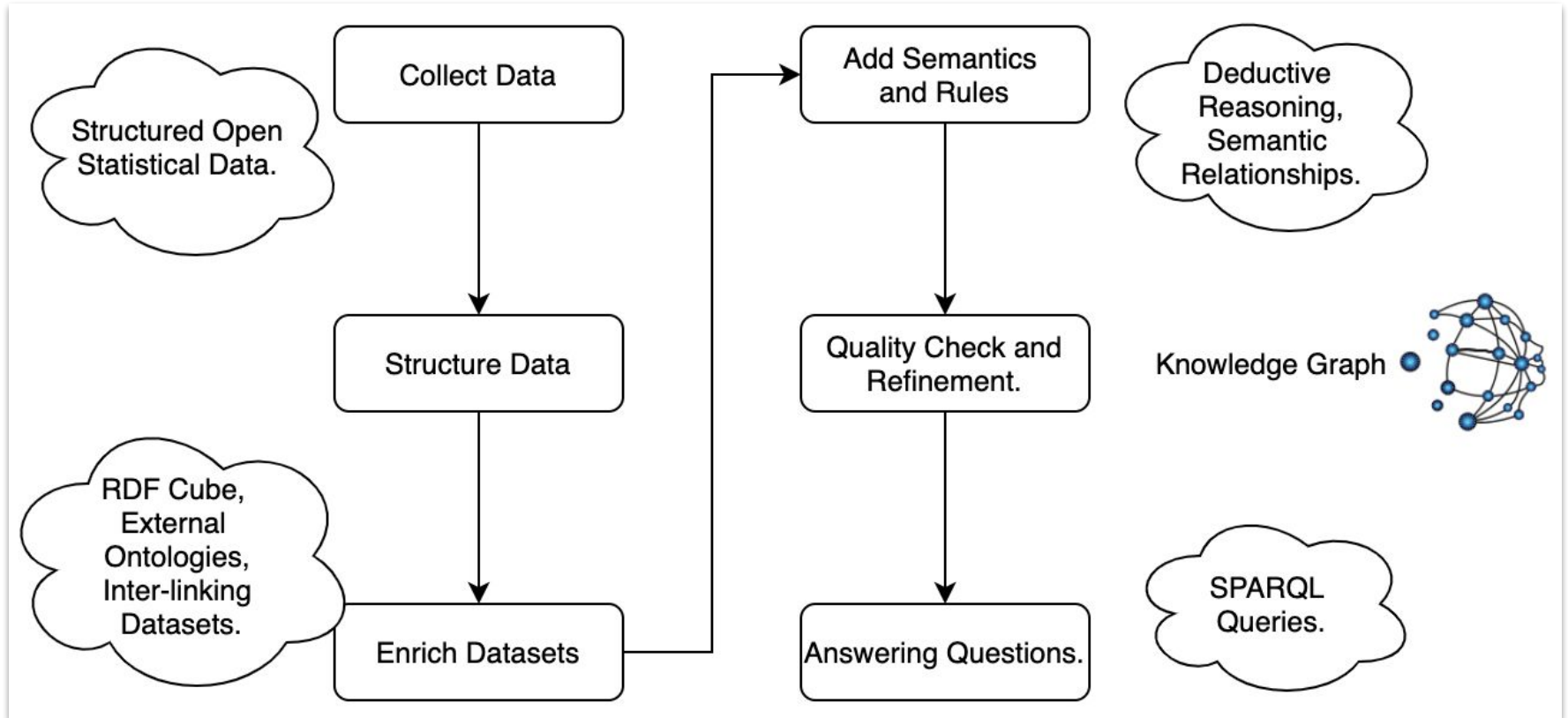
# Open Statistical Data: Observations

| Year | Disease                             | Number of Cases | Rate per 100,000 population |
|------|-------------------------------------|-----------------|-----------------------------|
| 2005 | Acquired Immune Deficiency Syndrome | 5               | 0.5                         |
| 2006 | Acquired Immune Deficiency Syndrome | 13              | 1.4                         |
| 2007 | Acquired Immune Deficiency Syndrome | 5               | 0.5                         |
| 2008 | Acquired Immune Deficiency Syndrome | 6               | 0.6                         |
| 2009 | Acquired Immune Deficiency Syndrome | 2               | 0.2                         |
| 2010 | Acquired Immune Deficiency Syndrome | 5               | 0.5                         |
| 2011 | Acquired Immune Deficiency Syndrome | 4               | 0.4                         |
| 2012 | Acquired Immune Deficiency Syndrome | 2               | 0.2                         |
| 2013 | Acquired Immune Deficiency Syndrome | 0               | 0                           |
| 2014 | Acquired Immune Deficiency Syndrome | 2               | 0.2                         |
| 2005 | Hepatitis B - Acute                 | 10              | 1.1                         |
| 2006 | Hepatitis B - Acute                 | 8               | 0.9                         |
| 2007 | Hepatitis B - Acute                 | 9               | 1                           |

< Previous   Next >

Showing Rows 1 to 13 out of 689

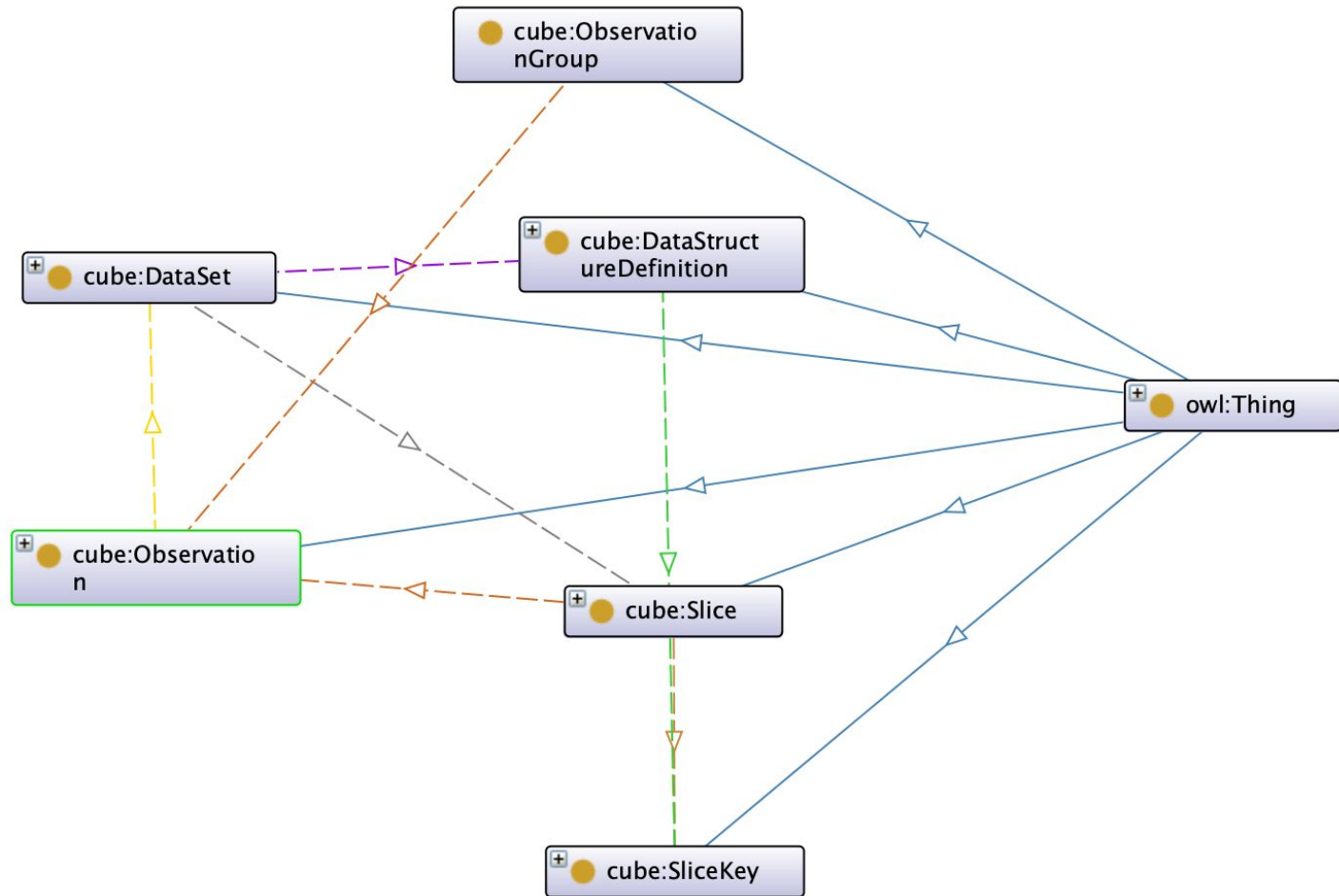
# Knowledge Graph Construction Process



# Creating Knowledge Graph for Disease Datasets

- Data collection: Used Python to download Nova Scotia datasets via Socrata API
- Data preparation: Analyzed the dataset and found 21 disease datasets
- Ontology: Created an ontology to struct the disease datasets
- Enrichment: Connected the datasets to the disease ontology (<https://disease-ontology.org/>)
- Semantic rules: Created SWRL rules in the knowledge graph to answer queries
- Published the constructed RDF knowledge graph on Zenodo website <https://zenodo.org/record/5517236#.YmuvzvPML6Y>

# Ontology based on RDF Cube Library



# Query Answering

- What are the viral infectious diseases in Nova Scotia?

```
SELECT ?disease_label ?disease_parent ?numofcases ?year
{
  ?observation eg:hasdisease ?disease.
  ?observation rdfs:label ?disease_label.
  ?disease rdfs:label ?disease_parent.
  FILTER regex(?disease_name,
               "viral infectious disease", "i")
  ?observation eg:numberofcases ?numofcases.
  ?observation dimension:refPeriod ?year.
}
```

# Conclusion and Next Steps

- Constructed a knowledge graph for open statistical data using the Semantic Web technologies and tools (RDF, SWRL, Protégé).
- Connected 21 statistical datasets in disease domain.
- The collected datasets had same structure. However, this is not the case in the other datasets.
- There are two challenges:
  - How to connect different datasets with different structures?
  - How to connect similar datasets in different provinces?